Risikoer ved maskinlæringsalgoritmer

Jan Roar Beckstrøm
avdelingsdirektør/chief data scientist
Riksrevisjonens innovasjonslab

For ordens skyld: Dette foredraget er mine synspunkter, og gjenspeiler ikke nødvendigvis Riksrevisjonens syn

Maskinlæring & KI er en god ting

- Kan bidra til vesentlig forbedret produktivitet i offentlig og privat sektor
- Kan gi oss fantastiske nye tjenester (Fintech?)
- ML er «overalt», men vi har likevel så vidt startet
- Vil eksempelvis ikke ha råd til en offentlig sektor som ikke bruker ML/KI i utstrakt grad

Dog, med stor oppside følger også stor risiko…

Support the Guardian
Available for everyone, funded by readers
Contribute → Subscribe →

Search jobs | Sign in | Search | The Guardian | International edition

News | Opinion | Sport | Culture | Lifestyle | More

Coronavirus World UK Environment Science Global development Football Tech Business Obituaries

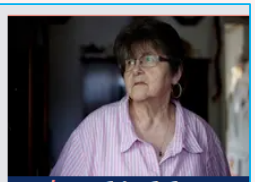
Automating poverty

A series exploring how our governments use AI to target the vulnerable

Digital dystopia / How algorithms punish the poor

Comment
Benefits of 'welfare robots' and the need for human oversight

UN
Big tech allowed to target and surveil the poor creating 'digital welfare state'



US / Zombie debts are hounding struggling Americans. Will you be next?

Computer says no / The people trapped in universal credit's 'black hole'

How Bristol assesses citizens' risk of harm – using an algorithm

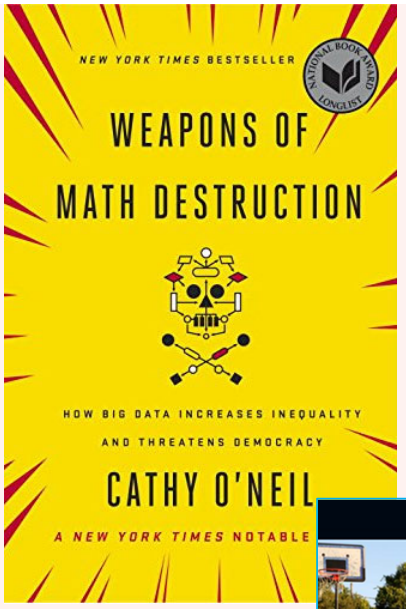
Around the world

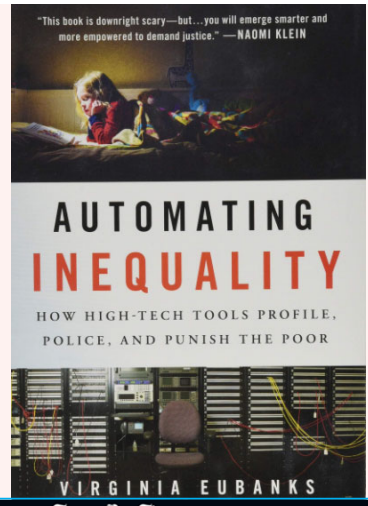
India / How a glitch in biometric welfare system can be lethal

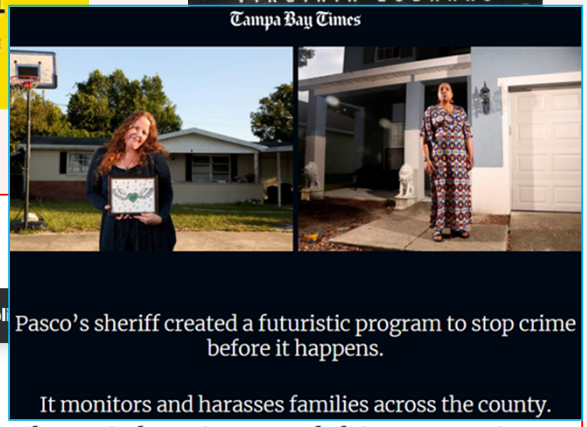
Australia / The automated system leaving welfare recipients cut off with nowhere to turn

UK / Benefits system automation could plunge claimants deeper into poverty

The Guardian view / On automating poverty: OK computers?







diginomica

Core tech | Future tech | Pol

UN report - our algorithmic world is creating a social welfare dystopia

By Jerry Bowles October 22, 2019 6 min reading

SUMMARY: Tech marketers dish out plenty of excitement about intelligent software and automating the inefficient. But a new UN report raises concerning questions about our algorithmic futures.

https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25156

Predictive policing

LOS ANGELES POLICE COMMISSION

REVIEW OF SELECTED
LOS ANGELES POLICE DEPARTMENT
DATA-DRIVEN POLICING STRATEGIES

Conducted by the

OFFICE OF THE INSPECTOR GENERAL

MARK P. SMITH
Inspector General

March 12, 2019

http://www.lapdpolicecom.lacity.org/031219/BPC_19-0072.pdf

The Atlantic

DO ALGORITHMS HAVE A PLACE IN POLICING?

How a Pakistani-born retired pilot took on a controversial, data-driven policing program in Los Angeles—and won

By Eva Ruth Moravec

SEPTEMBER 5, 2019 SHARE

https://www.theatlantic.com/politics/archive/2019/09/do-algorithms-have-place-policing/596851/

Los Angeles Times

CALIFORNIA

LAPD will end controversial program that aimed to predict where crimes would occur

In 2019, Officers Denise Vasquez and Oscar Bocanegra patrol a Tarzana area where a computer program predicted a higher possibility of property crime. (Mel Melcon / Los Angeles Times)

BY LEILA MILLER | STAFF WRITER
APRIL 21, 2020 UPDATED 6:17 PM PT

SUBSCRIBERS ARE READING

MUSIC
For Travis Scott, a history of chaos at con followed by a night of unspeakable traged

LIFESTYLE
The L.A. Times 2021 holiday gift guide

CALIFORNIA
FOR SUBSCRIBERS
They fled L.A. for Joshua Tree during the pandemic. Now they face the reality of de life

OPINION
Op-Ed: As a USC professor, I can't stay qu about the administration's toxic culture

CALIFORNIA
Faced with soaring Ds and Fs, schools are ditching the old way of grading

https://www.latimes.com/california/story/2020-04-21/lapd-ends-predictive-policing-program

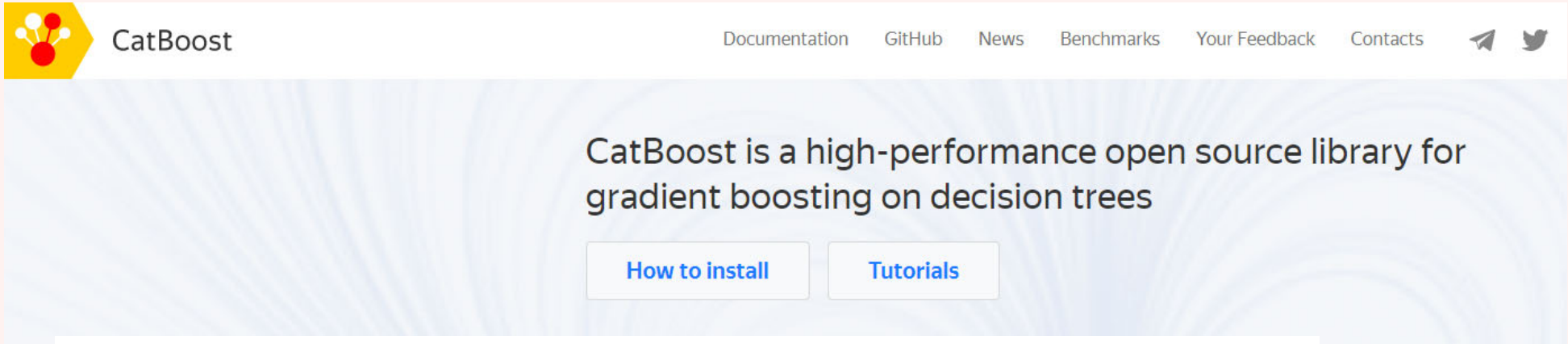
Robotvedtak i offentlig forvaltning

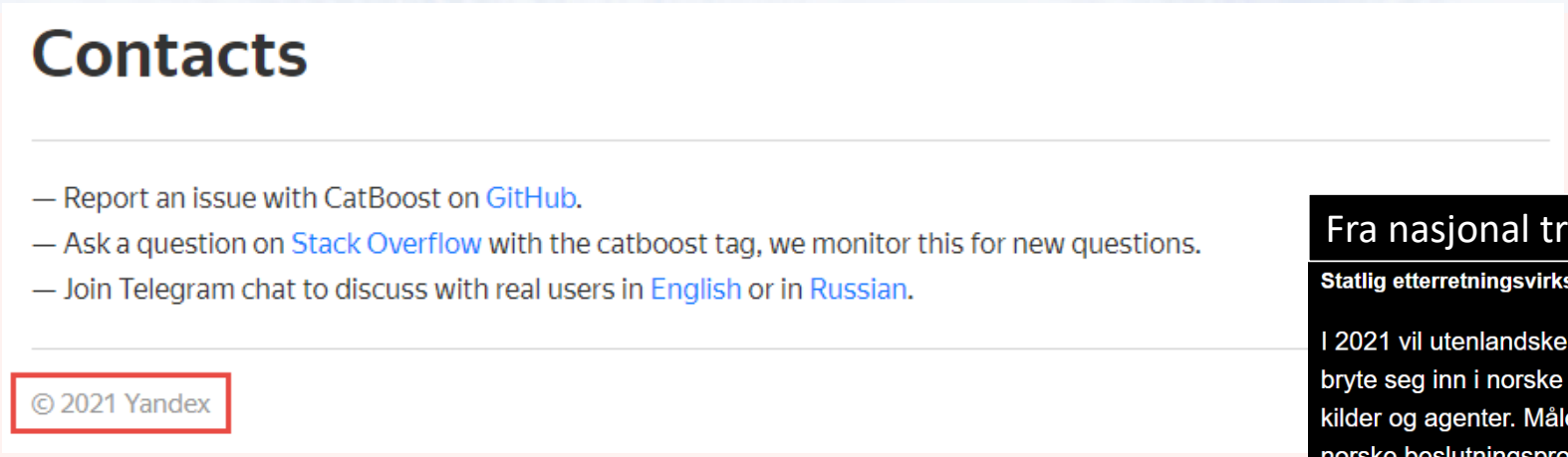
– rettssikkerhetsutfordringer ved bruk av helautomatiserte saksbehandlingssystemer

Sitat: «Risikoen for masseproduksjon av feil er stor, og det er derfor grunn til å stille strenge krav til testing av systemet i forkant, og hyppig systemkontroll etter at systemet er tatt i bruk.»

ML-algos: Hvor kommer de fra…

https://catboost.ai/





Fra nasjonal trusselvurdering 2021:

Statlig etterretningsvirksomhet

I 2021 vil utenlandske etterretningstjenester bruke store ressurser på å bryte seg inn i norske datanettverk. De vil også forsøke å rekruttere kilder og agenter. Målet deres er å få tilgang til informasjon og å påvirke norske beslutningsprosesser. Russisk og kinesisk etterretningsaktivitet vil utgjøre den største trusselen.

Samtidig…

Det kan tross alt hende maskinelt «skjønn» kan lages bedre enn menneskelig skjønn?

Sitat:

“We are still using these algorithms called humans that are really biased. We’ve tested them and known that they’re horrible, but we still use them to make really important decisions every day.”

(Rayid Ghani, computer scientist, Carnegie Mellon University)

Hva om maskiner hadde gjort stort sett all saksbehandling i både privat og offentlig sektor?

Hvor mye kunne vi spare på det?

Nytte > (menneskelige) kostnader?

Men…

- Er ikke helautomatisert saksbehandling forbudt etter § 22 i GDPR?

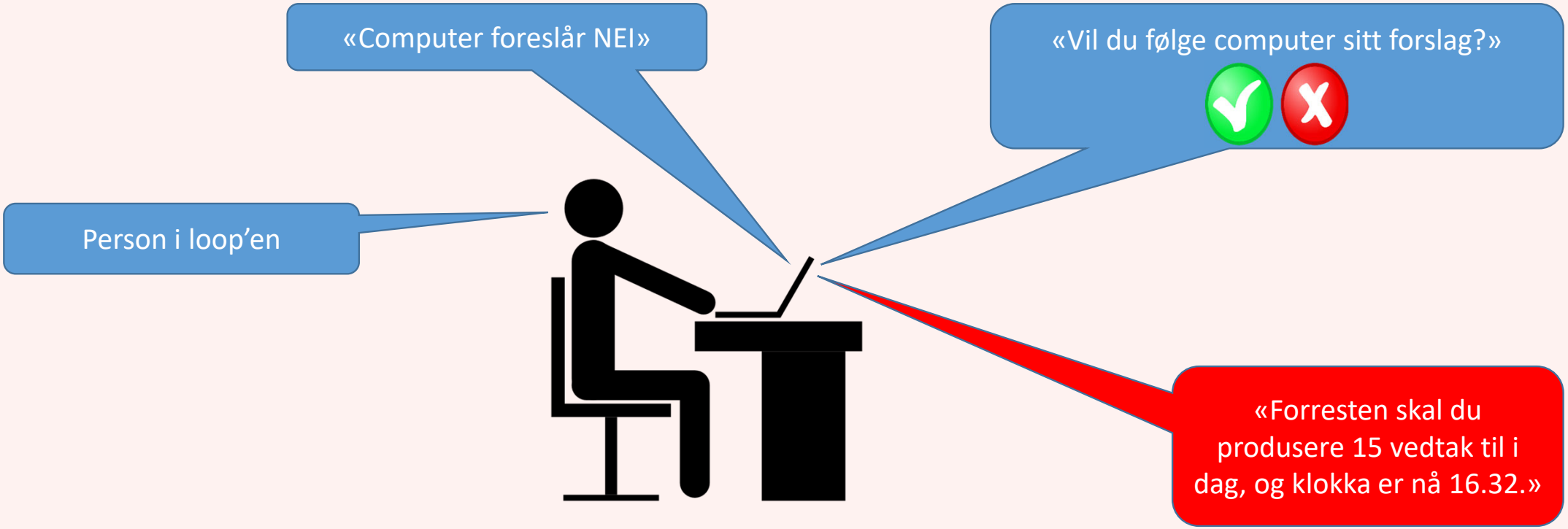
- GDPR Artikkel 22 “Automated individual decision-making, including profiling“ sier som følger:

“The data subject shall have the right not to be subject to a decision based solely on automated processing”

Kort sagt: Det skal være **“en person i loop’en”**

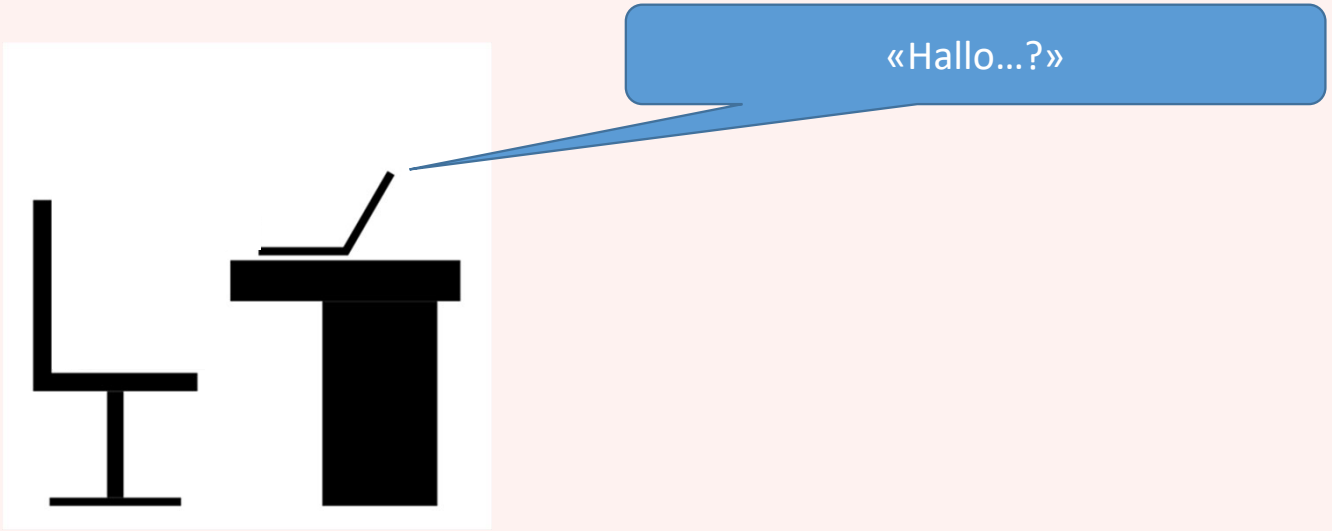
Automatisert, dystopisk saksbehandling…

(Beslutningsstøttesystem basert på maskinlæring - HAL 9000)



Personen ble sykemeldt…

(Beslutningsstøttesystemet basert på maskinlæring føler seg forlatt)



Problematiske sider med ML-modeller – to eksempler

Black box: ML-algoritmer er vanligvis sorte bokser, og det er uklart hvordan de kommer fram til et gitt utfall/resultat

Bias: ML-modeller vil plukke opp bias fra treningsdata. Dette kan gi systematisk feil resultater i form av diskriminering og ubegrunnet forskjellsbehandling

Kilde:
https://christophm.github.io/interpretable-ml-book/

Black box – høyst forenklet

- Om du bruker en black box algoritme som f eks 

- for å avgjøre hvem som skal få/ikke få…
- en bestemt stønad:

- Da er det faktisk umulig å vite sikkert…
- hvorfor en bestemt person…
- fikk/ikke fikk stønaden.

White box

I white box modeller er det mulig å si hvilken betydning en variabel (f eks. kjønn, alder, etnisitet osv.) har på personnivå

Således også bedre enn menneskelig skjønn?

F eks logistisk regresjon:
Du får et tall på hvor sannsynlig det er at utfall A (og ikke B) er riktig, for enkeltpersoner

Modellen er forklarbar

Bias – et eksempel

nature

View all journals | Search | Login

Explore content | Journal information | Publish with us | Subscribe | Sign up for alerts | RSS feed

nature › news › article

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford

You have full access to this article via your institution.

Download PDF

Related Articles

A fairer way forward for AI in health care

Bias detectives: the researchers striving to make algorithms fair

"Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients."

https://www.nature.com/articles/d41586-019-03228-6

Maskinell diskriminering og loven…

Kapittel 2 Forbud mot å diskriminere

§ 6. *Forbud mot å diskriminere*

Diskriminering på grunn av kjønn, graviditet, permisjon ved fødsel eller adopsjon, omsorgsoppgaver, etnisitet, religion, livssyn, funksjonsnedsettelse, seksuell orientering, kjønnsidentitet, kjønnsuttrykk, alder eller kombinasjoner av disse grunnlagene er forbudt. Med etnisitet menes blant annet nasjonal opprinnelse, avstamning, hudfarge og språk.

Black box ←→ Bias – en sammenheng

I black box modeller vil vi ikke vite på personnivå hvorvidt slike faktorer, eller kombinasjon av faktorer blir overdrevet

Det er ofte en mulighet for at en modell plukker opp et signal («Aha! Kjønn er viktig!»), og overdriver betydningen av f eks kjønn

Ikke alltid enkelt å vite om resultatet er «biased» ettersom du ikke vet hva som skjer inni modellen

Riksrevisjonen

Saksnr: 2019/01316

Unntatt offentlighet jf. Offentleglova § 14 første ledd

Rapport

Bokontroll basert på maskinlæring

- En vurdering ved Riksrevisjonen i samarbeid med Statens lånekasse for utdanning

Det er ingenting i dokumentasjonen vi har fått som tilsier at Lånekassen har vurdert likebehandling i utviklingen av ML-modellen. Det kunne nok med fordel vært gjort, særlig ettersom det er mulig å belyse ved enkle analyser. Vi har derfor gjort noen slike analyser, som viser at modellen overdriver relativt kraftig betydningen av nettopp kjønn. Det ser derfor ut til at menn i større grad blir plukket ut som «sannsynlige misligholdere» enn det det er grunnlag for i datamaterialet. Den reelle andelen som har oppgitt feil bostatus[36] er 4 % kvinner og 7 % menn. Det er derfor gode grunner til at menn vil ha noe høyere sannsynlighet for å bli plukket ut til kontroll enn kvinner. Imidlertid er FRP-verdien[37] 2,27 ganger så stor for menn sammenlignet med kvinner. Det betyr at menn som faktisk kan

[36] Basert på treningsdata

[37] "False positive rate", se appendiks A2

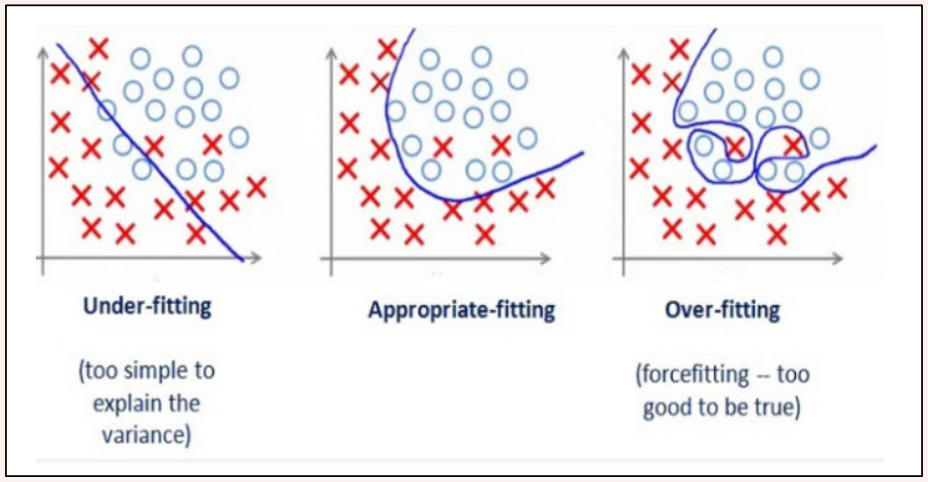
Konkrete risikoer – et par eksempler

- Overtrening
- Forgiftning av data (data poisoning)

Overtrening
(a.k.a. «overfitting»)

- Overtrening – et kronisk problem ved veiledet ML

- Forenklet:
  - Du lager (trener) en klassifiseringsmodell på basis av kjente (trenings-) data
  - For å predikere klassetilhørighet for nye enheter med ukjente data
  - Alltid en risiko for at modellen passer «for godt» til treningsdata → resulterer i dårlig prediksjonskraft



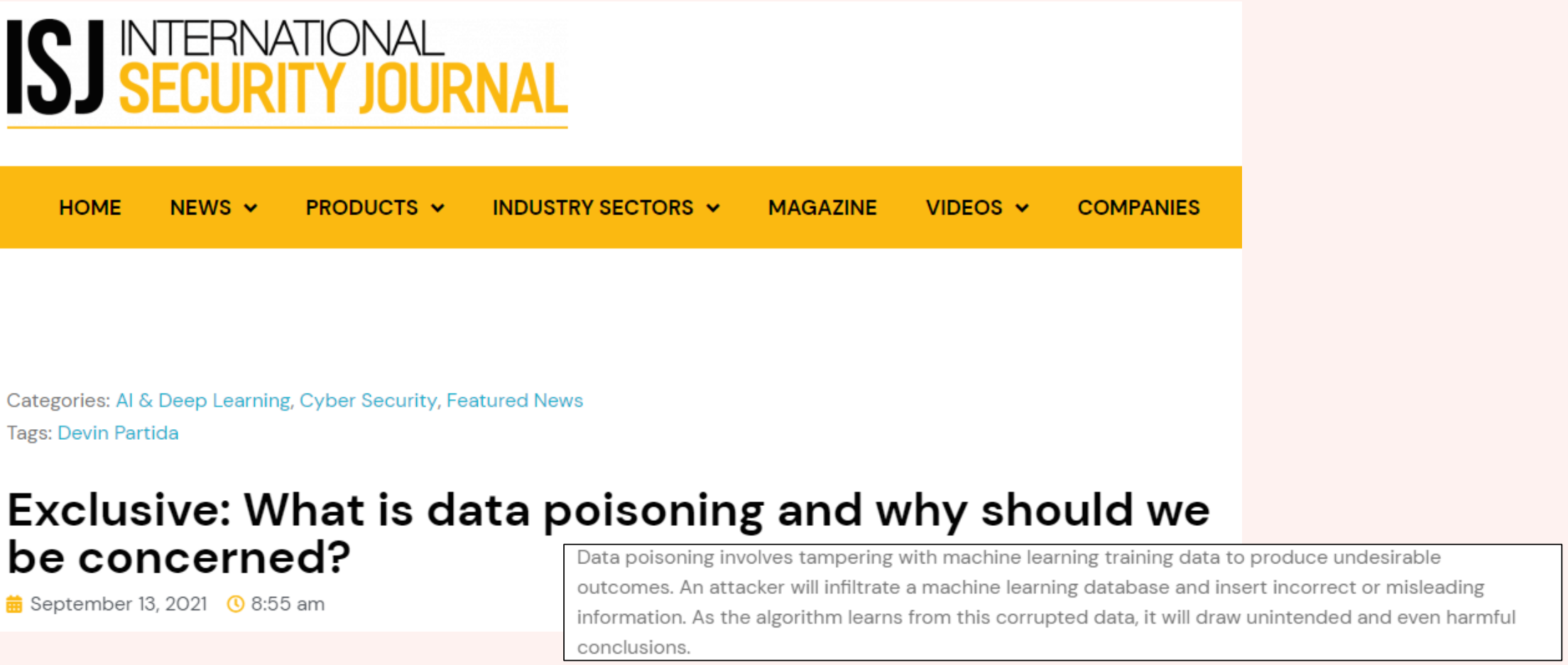
Source: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Overtrening – sagt på en annen måte…

The “Wayne Rooney” effect

- One way to notice overtraining is by time effects.
  - Time changes public opinion on particular people or effects.
  - Vampire movies go out of fashion, superhero movies come into fashion.
  - People who were hailed as superstars in 2003 might later get bad press in 2010
  - Called the “Wayne Rooney” effect

Forgiftning av data



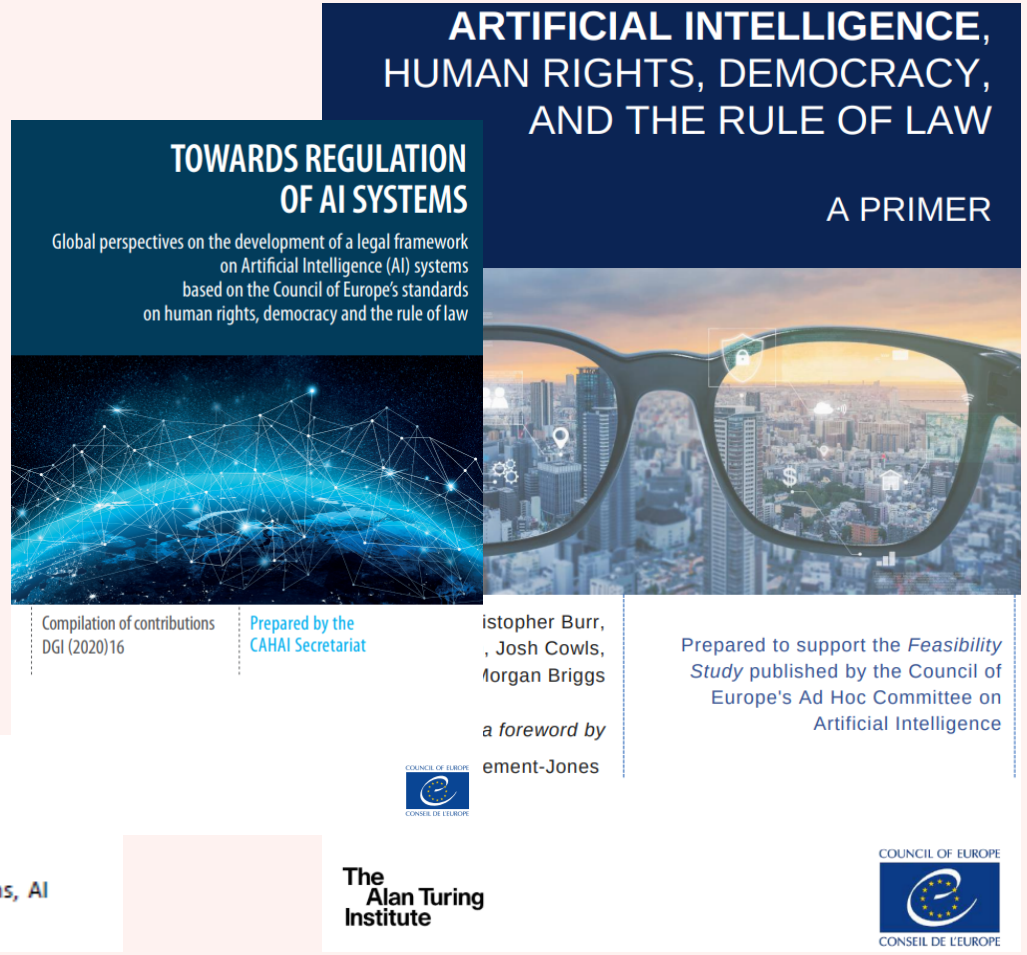
Kilde: https://internationalsecurityjournal.com/what-is-data-poisoning/

Håndtering av risiko – regulering og revisjon

Regulering av AI

Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

1. it should be lawful, complying with all applicable laws and regulations;
2. it should be ethical, ensuring adherence to ethical principles and values; and
3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.



Under the authority of the Committee of Ministers, the CAHAI is instructed to:

- examine the feasibility and potential elements on the basis of broad multi-stakeholder consultations, of a legal framework for the development, design and application of artificial intelligence, based on the Council of Europe's standards on human rights, democracy and the rule of law.

…og selvsagt EUs nye forslag + GDPR

EUROPEAN
COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

EUs nye AI regulering – noen highlights

Article 64
Access to data and documentation

1. Access to data and documentation in the context of their activities, the market surveillance authorities shall be granted full access to the training, validation and testing datasets used by the provider, including through application programming interfaces ('API') or other appropriate technical means and tools enabling remote access.

2. Where necessary to assess the conformity of the high-risk AI system with the requirements set out in Title III, Chapter 2 and upon a reasoned request, the market surveillance authorities shall be granted access to the source code of the AI system.

Article 10
Data and data governance

1. High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5.

2. Training, validation and testing data sets shall be subject to appropriate data ...nent practices. Those practices shall concern in particular,

... choices;

...ration processing operations, such as annotation, labelling, ...nt and aggregation;

...of relevant assumptions, notably with respect to the ...e data are supposed to measure and represent;

... of the availability, quantity and suitability of the data sets

...w of possible biases;

(g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.

3. Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.

NAV

Sluttrapport fra sandkasseprosjektet med NAV

Temaer: rettslig grunnlag, rettferdighet og forklarbarhet

Januar 2022

1. Sammendrag

Mål med sandkasseprosjektet

NAV ønsker å bruke maskinlæring til å forutse hvilke sykmeldte brukere som vil ha behov for oppfølging to måneder frem i tid. Dette skal hjelpe veilederne med gjøre mer treffsikre vurderinger, som igjen skal spare NAV, arbeidsgivere og de sykmeldte for unødvendige møter. Målet med dette sandkasseprosjektet var å avklare lovligheten ved bruk av kunstig intelligens (KI) i denne sammenhengen, og utforske hvordan profileringen av sykmeldte kan gjøres på en rettferdig og åpen måte.

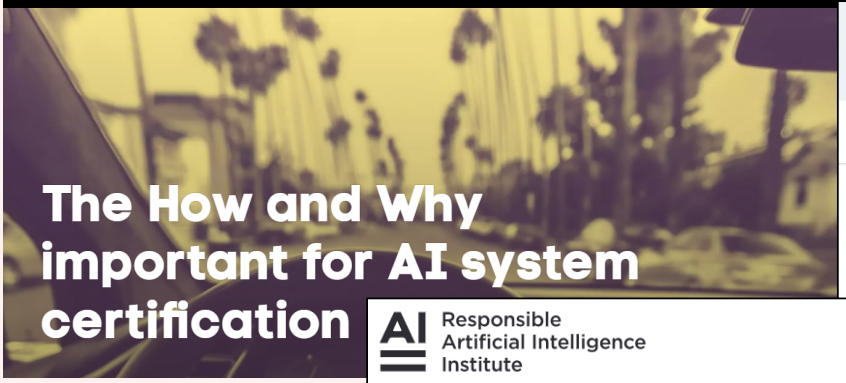
Konklusjoner

- **Lovlighet.** NAV har rettslig grunnlag for å bruke KI som støtte ved beslutning om enkeltindividers behov for oppfølging og dialogmøte. Det er usikkert om det rettslige grunnlaget åpner for å bruke personopplysninger til å utvikle selve algoritmen.
- **Rettferdighet.** Det er viktig forskjell mellom å benytte opplysninger som allerede inngår i modellen, og å ta i bruk nye opplysninger som ikke brukes i modellen, til å sjekke for diskriminerende utfall. Det oppstår en spenning mellom personvern og rettferdighet når metoden for å avdekke og motvirke diskriminering fordrer mer behandling av personopplysninger.
- **Åpenhet**. For at modellen skal gi ønsket verdi, er det avgjørende at NAV-veilederne stoler på algoritmen. Innsikt og forståelse i modellens virkemåte er viktig for å vurdere prediksjonen på et selvstendig og trygt grunnlag, uavhengig av om den endelige avgjørelsen blir å følge prediksjonens anbefaling eller ikke.

Sertifisering av ML-systemer blir en greie

RI.SE Our offer Industry About RISE Work with us Press Log in Search EN

Home → Our stories → The How and Why important for AI system certification Share:



The How and Why important for AI system certification

TÜV SÜD INDUSTRIES & SERVICES RESOURCES

SUBSCRIBE FOR INFORMATION

TÜV SÜD AND DFKI DEVELOP "TÜV FOR ARTIFICIAL INTELLIGENCE"

The German Research Center for Artificial Intelligence (DFKI) and TÜV SÜD are launching a joint project to certify systems based on artificial intelligence (AI) used in autonomous driving and develop a 'roadworthiness test' for algorithms. To do so, the experts will explore the learning behaviours of AI systems with the aim of being able to control the systems' reactions.

AI Responsible Artificial Intelligence Institute

RAII Certification Beta

The world's first independent, accredited certification program of its kind.

Developed under the Global AI Action Alliance for the World Economic Forum (WEF), along with a diverse community of leading experts, RAII Certification is based on objective assessments of fairness, bias, explainability, and other concrete metrics of responsibly built AI systems.

The EESC proposes introducing EU certification for "trusted AI" products

This page is also available in fr

14/11/2019

Related content See also

The European Economic and Social Committee (EESC) suggests that the EU should develop a certification for trustworthy AI applications, to be delivered by an independent body after testing the products for key requirements such as resilience, safety, and absence of prejudice, discrimination or bias. The proposal has been put forward in two recent EESC opinions assessing the European Commission's ethical guidelines on AI.

Both EESC opinions – one covering the communication on *Building trust in human-centric artificial intelligence* as a whole and the other on its specific implications for the automotive sector – stress that such a certification would go a long way towards increasing public trust in artificial intelligence (AI) in Europe.

…og ikke minst revisjon av algoritmer…

www.auditingalgorithms.net

- En guide skrevet av revisorer (oss), for revisorer
- et internasjonalt samarbeid (riksrevisjonene i Tyskland, Storbritannia, Nederland, Finland, Norge)
- Stort sett ikke-teknisk